Statistical Analysis of Clustered Data using SAS® System

Gui-shuang Ying, Ph.D.
Chengcheng Liu, M.S.
Center for Preventive Ophthalmology and Biostatistics,
Department of Ophthalmology, University of Pennsylvania

ABSTRACT

Clustered data is very common, such as the data from paired eyes of the same patient, from multiple teeth of the same mouth, from animals of the same litter, from siblings in the same family. The key feature of clustered data is that outcomes from the same cluster are likely to be positively correlated. The proper analysis of clustered data requires taking this correlation into consideration. The ignorance of such correlation can bias the statistical inference.

This paper provides an overview on the availability of built-in SAS procedures and user-developed SAS macros for the analysis of clustered data from Ophthalmology studies. It covers the non-model based analysis by PROC TTEST, PROC UNIVARIATE and %CLUSWILCOX for the parametric and nonparametric comparison of paired continuous data; PROC FREQ, %MHADJUST and %CLUSTPRO for the comparison of balance or unbalanced paired binary data; followed by the model-based analysis using PROC GENMOD, PROC MIXED, PROC GLIMMIX, PROC NLMIXED, PROC PHREG and %GAMFRAIL for clustered continuous, binary, count and survival data. We conclude that SAS is very powerful for analyzing clustered data.

INTRODUCTION

Clustered data arises from many applications, including ophthalmologic studies, rodent teratology experiments, dental research, family-based genetic studies, and community intervention studies, etc. In ophthalmology studies, it is common to randomize one eye of each subject to the treatment and the other eye as control (randomization unit is eye), or randomize paired eyes of the same subject into the same treatment (randomization unit is patient), and eye specific measurements such as visual acuity (VA), refraction error, intra ocular pressure (IOP), cataract status etc. from both eyes are obtained. The measurements from the paired eyes of same subject tend to be positively correlated, due to the common subject-specific characteristics such as age, diet, and genetic factors.

In the analysis of such clustered data, estimates of effect (such as mean differences, odds ratios) might be accurately derived from clustered data without adjusting correlation; however, the variability of these effects would likely be biased, leading to incorrect test statistics and confidence intervals. For example, if correlation from paired eyes was ignored, the standard error for the treatment effect is likely to be overestimated when paired eyes from the same subject are in different treatment group, while it is likely to be underestimated when paired eyes are in the same treatment group. For this reason, most statistical techniques, such as the unpaired t-test for comparison of means, or chi-square test for comparison uncorrelated proportions will not work because they assume that observations from the same cluster are independent. The appropriate statistical analysis of such clustered data needs to take correlation into consideration, otherwise the results obtained will not be valid.

This paper describes the available built-in SAS procedures and user-developed SAS macros to analyze clustered data in general, with data from Ophthalmology studies in particular. It describes the simple non-model based analysis by PROC TTEST, PROC UNIVARIATE, and %CLUSWILCOX for the parametric and nonparametric analysis of paired continuous data; PROC FREQ, %MHADJUST and %CLUSTPRO for the comparing of correlated proportions of balanced and unbalanced paired data. Examples are provided for the model-based analysis using PROC GENMOD, PROC MIXED, PROC GLIMMIX, PROC NLMIXED for clustered continuous, binary, count and ordinal data; PROC PHREG and frailty models using SAS macros for clustered time to event data.

We demonstrate these analyses using the data from the Bilateral Drusen Study of the Chroidal Neovascularization Prevention Trial (CNVPT). In this trial, 156 patients with both eyes showing high-risk nonexudative age-related macular degeneration (AMD) were enrolled, with one eye randomly assigned to laser treatment, the other eye as control, and their visual acuity (VA) was measured at baseline, 6 months, and annually for four years after treatment (CNVPT Research Group, 1998).

ANALYSIS OF CLUSTERED CONTINUOUS DATA

When the outcome is continuous, such as VA measured as number of letters read correctly from ETDRS charts (it ranges from 0 to 95, higher value indicates better visual acuity), the paired t-test using PROC TTEST for the paired data could be performed (Figure 1, 2); or equivalently, the one-sample t-test could be performed for the derived difference between paired eyes by using PROC UNIVARIATE (Figure 3). When nonparametric test is needed due to non-normality, the signed rank test could be used for the difference by PROC UNIVARIATE (Figure 3).

Fig. 1 SAS Code for comparing VA at 48 months between treated vs. observed eyes

```
data laser(keep=id va48 vachg48 loss3 rename=(va48=l_va48 vachg=l_vachg48 loss3=l_loss3))
      observed(keep=id va48 vachg48 loss3 rename=(va48=o_va48 vachg=o_vachg48 loss3=o_loss3));
   set bdata;
   if group=1 then output laser;
   else if group=0 then output observed;
run;

proc sort data=laser; by id;
proc sort data=observed; by id;

data botheye;
  merge laser observed; by id;
   vadiff48=l_va48-o_va48;
run;

proc ttest data=botheye;
   paired l_va48*o_va48;
run;

proc univariate data=botheye;
   var vadiff48;
run;
```

Fig. 2 SAS Output from Paired ttest

The TTEST Procedure

Statistics

Difference	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev
l_va48 - o_va48	98	-3.877	0.4796	4.8364	19.056	21.731	25.286

Statistics

Difference	Std Err	Minimum	Maximum
l_va48 - o_va48	2.1952	-65	85

T-Tests

Difference	DF	t Value	Pr > |t|
l_va48 - o_va48	97	0.22	**0.8275**

Fig. 3 SAS Output from One-sample ttest and signed rank test

The UNIVARIATE Procedure
Variable: vadiff48

Moments

N	98	Sum Weights	98
Mean	0.47959184	Sum Observations	47
Std Deviation	21.730889	Variance	472.231538
Skewness	0.5688634	Kurtosis	4.29671576
Uncorrected SS	45829	Corrected SS	45806.4592
Coeff Variation	4531.12154	Std Error Mean	2.19515128

Basic Statistical Measures

Location		Variability	
Mean	0.479592	Std Deviation	21.73089
Median	0.000000	Variance	472.23154
Mode	0.000000	Range	150.00000
		Interquartile Range	13.00000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	0.218478	Pr > |t|	**0.8275**
Sign	M	3	Pr >= |M|	0.5984
Signed Rank	S	125.5	Pr >= |S|	**0.6161**

When all the subunits of a cluster are in the same comparison group, and the data is not normally distributed, the nonparametric Wilcoxon rank sum test that incorporates the cluster effects is needed (Rosner, 2003), this can be conducted by SAS macro %CLUSWILCOX for both balanced (same number of subunits per cluster) or unbalanced (different number of subunits per cluster) data. This macro could be found at http://www.tibs.org/biometrics/datasets/cluswilcox.sas.pdf. For demonstration purpose only, we compared baseline VA (from both eyes) between male and female patients using %CLUSWILCOX, the SAS output was shown in Figure 4.

Fig. 4 SAS Output from SAS macro CLUSWILCOX

Clustered Wilcoxon RankSum Statistic	Expected Value Clustered Wilcoxon RankSum Statistic	Variance of Clustered Wilcoxon RankSum Statistic	Z statistic for Clustered Wilcoxon RankSum Statistic	P-value for Clustered Wilcoxon RankSum Z Statisic
27958	29735	900934.25	-1.87215	0.061186

The above non-model based analyses are simple to implement and easy to understand, but the effect from other covariates could not be adjusted or studied. When we are interested in the eye-level or patient-level covariates, we have to use model based analysis. One commonly used model-based analysis of clustered data is to fit the marginal generalized estimating equations (GEE) regression models (Liang and Zeger, 1986) using PROC GENMOD. In GEE, the dependence within cluster is treated as nuisance, and random effects are not incorporated in the marginal model. The merit of GEE is that valid inferences are produced for population average effects as long as the mean structure is correctly specified, even if the dependence structure is misspecified. The application of GEE methodology using SAS has been detailed elsewhere (Johnston and Strokes, 1997).

The second model-based analysis for clustered continuous data is the mixed model using PROC MIXED (Littell, Milliken, Stroup, and Wolfinger, 1996). The mixed model incorporates both random and fixed effects into the model, it assumes that the random effects account for the correlation between measures from the same cluster.

Besides GEE and mixed models, we can also fit the clustered continuous data by generalized linear mixed model (GLMM) using new GLIMMIX procedure (GLIMMIX Procedure, 2005). This procedure was made available last August as an experimental procedure and as the production version this June, by download only, for the Windows platform and works with the SAS 9.1 release. The GLIMMIX procedure and its document could be found at http://support.sas.com/rnd/app/da/glimmix.html.

For demonstration, we performed model-based comparison of VA at 48 months between treated and observed eyes. The SAS codes for PROC GENMOD, PROC MIXED and PROC GLIMMIX are in Figure 5; SAS output are in Figure 6. The p-values from these three model-based analyses are extremely similar, they are also very similar to that from paired ttest.

Fig. 5 SAS Code for PROC MIXED, PROC GLIMMIX and PROC GENMOD

```
proc genmod data=bdata;
 class id;
 model va48=group/dist=normal;
 repeated sub=id/type=ind;
run;

proc mixed data=bdata;
 class id;
 model va48=group;
 repeated/sub=id type=cs;
run;

proc glimmix data=bdata;
 class id;
 model va48=group/dist=normal;
 random int/subject=id;
run;
```

Fig. 6 SAS Output from GENMOD, MIXED, and GLIMMIX

The GENMOD Procedure

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	74.9286	1.7336	71.5308	78.3263	43.22	<.0001
GROUP	0.4796	2.1839	-3.8008	4.7600	0.22	**0.8262**

The Mixed Procedure

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
GROUP	1	97	0.04	**0.8428**

The GLIMMIX Procedure

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
GROUP	1	97	0.05	**0.8275**

ANALYSIS OF CLUSTERED BINARY DATA

When the clustered data are binary, such as ≥ 3-lines loss in VA at 48 months (VA score decreased by ≥15 letters from baseline), the comparison of correlated proportions could be made by McNemar's test using PROC FREQ with AGREE option specified; or equivalently, using the Cochran-Mantel-Haenszel (CMH) test in PROC FREQ with CMH option specified (Figure 7, 8). However, when the paired data is not totally matched, for example, when the data from one eye is available while data from the other eye is missing, the McNemar's test excludes the un-matched pair data from analysis, leading to the loss of information. Duffy proposed a modified Mantel-Haenszel procedure that combines matched and unmatched binary data for analysis (Duff, Rohan,and Altman, 1989), this method could be easily performed in SAS through a little programming to assign the subjects with unmatched pair a common ID before applying the standard CMH procedure (Figure 9).

Fig. 7 SAS code for McNemar's test and CMH test

```
Proc freq data=botheye;
   tables l_loss3*o_loss3/agree;
run;

proc freq data=bdata;
   tables id*group*loss3/cmh noprint;
run;
```

Fig. 8 SAS Output for McNemar's test and CHM test

Statistics for Table of l_loss3 by o_loss3

McNemar's Test

Statistic (S)	0.3600
DF	1
Pr > S	**0.5485**

Summary Statistics for GROUP by loss3
Controlling for ID

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	0.3600	0.5485
2	Row Mean Scores Differ	1	0.3600	0.5485
3	General Association	1	0.3600	**0.5485**

Fig. 9 Modified CMH test that combine matched and unmatched binary data

```
data bdata2;
  set botheye;
   if l_loss3 ~=' ' and o_loss3 ~=' ' then id2=id;
   else id2=9999;  /* Assign a common ID for those with one eye data missing */
   loss3=l_loss3;
   trt="Laser";
   output;
   result=o_loss3;
   test="Observed";
   output;
run;

proc freq data=bdata2;
  tables id2*trt*loss3/cmh noprint;
run;
```

Obuchowski (Obuchowski, 1998) extended the McNemar's test to compare the correlated proportions for clustered data, where multiple glands from same patient was evaluated by two different diagnostic tests and comparison of sensitivity between these two tests was of interest. The extended McNemar's test can be executed by the SAS macro %CLUSTPRO (Lieber, 1998). This macro could be useful for the assessing inter-examiner difference when paired eyes are both examined by two different examiners, and could also be applied to analyze the data with two levels of correlation. For example, we applied this macro to compare the proportion of ≥3-lines loss in VA across 4-years longitudinal follow-up between laser and treated eyes, the SAS output is shown in the Figure 10.

Fig. 10 SAS Output from %CLUSTPRO

```
p_hat1 = 0.1096938776, p_hat2 = 0.1173469388, p_bar = 0.1135204082
chi-square statistic = 0.06502308911035
degrees of freedom = 1
p-value = 0.7987259932
S
0.000500707528   0.000084113766
0.000084113765   0.000568266900
```

For assessing the association between a binary outcome and a binary exposure while adjusting for a categorical covariate, the CMH procedure is commonly used. However, when the data are correlated in clustered data, the modified Mantel-Haenszel procedure (Begg, 1999) that adjusts for the correlation is needed. The SAS macro %MHADJUST that performs modified Mantel-Haenszel procedure could be found at http://www.columbia.edu/~mdb3/mhadjust.txt and more details are available in a published paper (Begg, Paykin, 2001).

The advantage of McNemar's test, extended McNemar's test or modified CMH procedure for the comparison of correlated proportion is its simplicity, and no particular correlation structure is assumed. However, the drawback is that the effect of covariates could not be examined. When the effects of covariates are of interest, we could use the model-based analysis by GEE, GLMM or nonlinear mixed model. Compared with non-model based analysis, the GEE or GLMM approach is more complicated in terms of both the model and computations, and needs the specification of working correlation structures. Besides, the nonlinear mixed model needs to provide initial values of parameters, and write your own model equation and variance-covariance structure. Additionally, we could use the conditional logistic regression using PROC LOGISTIC with STRATA option specified.

The following example shows the application of PROC GENMOD, PROC GLIMMIX, PROC NLMIXED and PROC LOGISTIC to compare the proportion of ≥3-lines loss in VA at 48 months between treated and control eyes. The SAS codes are in Figure 11, and SAS output in Figure 12. As shown in the Figure 12, the estimate of treatment effect and p-values from these four analyses are very similar.

Fig. 11 SAS Code for GENMOD, GLIMMIX, NLMIXED and LOGISTIC

```
proc genmod data=bdata descending;
   class id;
   model loss3=group/dist=bin link=logit;
   repeated sub=id/type=ind;
run;

proc glimmix data=bdata;
   class id;
   model loss3 (descening)=group/dist=bin link=logit cl;
   random int/subject=id;
run;

proc nlmixed data=bdata;
   parms b0=0, b1=0, sd=1;
   z=b0+b1*group+U;
   model loss3 ~binary(1/(1+exp(-z)));
   random U ~ normal(0, sd*sd) subject=id;
run;

proc logistic data=bdata descending;
   model loss3=group;
   strata id;
run;
```

Fig. 12 SAS Output from GENMOD, GLIMMIX and NLMIXED

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-1.6341	0.2733	-2.1698	-1.0985	-5.98	<.0001
GROUP	**0.2091**	**0.3485**	-0.4740	0.8922	0.60	**0.5485**

The GLIMMIX Procedure

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha
Intercept	-1.6518	0.2829	97	-5.84	<.0001	0.05
GROUP	**0.2110**	**0.3758**	97	0.56	**0.5758**	0.05

The NLMIXED Procedure
Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
b0	-1.8962	0.3931	97	-4.82	<.0001	0.05	-2.6763	-1.1160
b1	**0.2368**	**0.3991**	97	0.59	**0.5544**	0.05	-0.5553	1.0283
sd	0.9275	0.4856	97	1.91	0.0591	0.05	-0.03631	1.8913

Conditional Logistic Analysis

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
GROUP	1	**0.2412**	**0.4029**	0.3583	**0.5495**

ANALYSIS OF CLUSTERED COUNT DATA

We can model the clustered count data by Poisson regression. In SAS, it could be performed by GEE using PROC GENMOD or GLMM using PROC GLIMMIX. In these analyses, the options DIST = Poisson and LINK=log, should be specified, along with the correlation structure for the counts in the same cluster.

For demonstration purpose, we calculated the number of times that an eye having ≥3-lines loss in VA during four years of follow-up, and compared it between treated and control group by Poisson Model (Figure 13, 14).

Fig. 13 SAS Code for Poisson Regression Using GENMOD and GLIMMIX

```
proc genmod data=count descending;
   class id;
   model nloss3=group/dist=Poisson link=log;
   repeated sub=id/type=cs;
run;

proc glimmix data=count;
   class id;
   model nloss3=group/dist=Poisson link=log cl;
   random int/subject=id type=cs;
run;
```

Fig. 14 SAS Output for Poisson Regression Using GENMOD and GLIMMIX

The GENMOD Procedure

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-1.0014	0.2337	-1.4595	-0.5434	-4.29	<.0001
GROUP	**0.0541**	0.3196	-0.5723	0.6804	0.17	**0.8657**

The GLIMMIX Procedure

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept	-1.2780	0.2080	97	-6.14	<.0001	0.05	-1.6909	-0.8651
GROUP	**0.05407**	0.2326	97	0.23	**0.8167**	0.05	-0.4075	0.5157

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
GROUP	1	97	0.05	0.8167

When the count data is not well fitted by Poisson model, it could be analyzed using the negative binomial (NB) model through PROC GENMOD or PROC GLIMMIX by specifying options DIST =NB and LINK=log. We could also analyze the clustered count data using PROC NLMIXED, it can fit both Poisson and, starting from SAS version 8.1, NB model. Example 46.4 of the SAS/STAT User's Guide, Version 8 (1999) describes how to fit the Poisson model using PROC NLMIXED.

When the Poisson model or negative binomial model does not fit well with the count data, modifications to the Poisson model have been proposed to account for the over-dispersion by introducing a dispersion parameter. Poisson mode with over-dispersion parameter could be fit by using option DSCALE in the Model statement of PROC GENMOD as shown in Figure 15.

Fig. 15 SAS Code for the Poisson Regression with Over-dispersion

```
proc genmod data=count descending;
   class id;
   model nloss3=group/dist=Poisson link=log dscale;
   repeated sub=id/type=cs;
run;
```

When there is excessive zero in the count data, the zero-inflated Poisson model (Lambert, 1992) is more appropriate. This model assumes that the population is characterized by two regimes, one where members always have zero counts, and one where members have zero or positive counts. The likelihood of being in either regime is estimated using a logit specification, while the counts in the second regime are estimated using a Poisson specification. Zero-inflated model could be fitted by PROC TRAJ (Jones, Nagin, and Roeder, 2001), the document and procedure could be found in http://www.andrew.cmu.edu/user/bjones/index.htm.

ANALYSIS OF CLUSTERED ORDINAL DATA

For the ordinal data analysis, the proportional odds model is a popular method, which is based on modeling cumulative logit functions. In SAS, the clustered ordinal data could be analyzed using cumulative logistical regression by PROC GENMOD and PROC GLIMMIX, the multinomial distribution should be specified with the options DIST=mult, and link function specified as LINK=cumulative logit.

For the bivariate ordinal data, the Bivariate Dale model (BDM) could be fit (Dale, 1986) by the SAS macro %BDM (McMillam, and Hanson, 2005).

The cumulative logistic regression by PROC GENMOD and PROC GLIMMIX for VA change at 48 months (grouped into 3 ordinal lelves: worse, stable, better) is shown in Figure 16 and 17. Again, very similar results in both parameter estimate and P-values.

Fig. 16 SAS Code for Analysis of Ordinal Data using GENMOD and GLIMMIX

```
proc genmod data=bdata descending;
 class id;
 model chglevel=group/dist=mult link=cumlogit;
 repeated sub=id/type=ind;
run;

proc glimmix data=bdata;
 class id;
 model chglevel (descending)=group/dist=mult link=cumlogit cl;
 random intercept/subject=id;
run;
```

Fig. 17 SAS Output from GENMOD and GLIMMIX

The GENMOD Procedure

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept1	-2.3834	0.3040	-2.9793	-1.7875	-7.84	<.0001
Intercept2	0.9056	0.2206	0.4733	1.3379	4.11	<.0001
GROUP	0.1781	0.2903	-0.3910	0.7472	0.61	**0.5396**

The GLIMMIX Procedure

Solutions for Fixed Effects

Effect	chglevel	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept	1	-2.4007	0.2947	97	-8.15	<.0001	0.05	-2.9856	-1.8158
Intercept	0	0.9098	0.2163	97	4.21	<.0001	0.05	0.4806	1.3391
GROUP		0.1793	0.2926	96	0.61	**0.5416**	0.05	-0.4016	0.7601

ANALYSIS OF CLUSTERED TIME TO EVENT DATA

Methods for analyzing clustered time to event data are still under extensive development. The clustered time to event data was commonly fitted by marginal model using the option of COVSANDWICH in the PROC PHREG statement, and using the ID statement to indicate that observations with the same ID are from the same cluster. This procedure uses the robust sandwich estimate of Lin and Wei (Wei, Lin, and Weissfeld, 1989) for statistical inference.

The example code and SAS output from marginal model of time to first ≥3-lines loss in VA using robust sandwich estimate is shown in Figure 18 & 19. Of note, the variance from model based estimate that ignores the correlation is much larger than that from the robust sandwich estimate (0.426 vs. 0.181).

Fig. 18 SAS Code for Marginal Model of Clustered Survival Data

```
proc phreg data=vachg2 covs(aggregate) covm;
   model loss3vt*loss3(0)=group/ties=discrete rl;
   id id;
run;
```

Fig. 19 SAS Output from PROC PHREG Using Sandwich Variance Estimate

The PHREG Procedure

Analysis of Maximum Likelihood Estimates
with Model-Based Variance Estimate

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
GROUP	1	0.18036	**0.42561**	0.1796	0.6717

Analysis of Maximum Likelihood Estimates
with Model-Based Variance Estimate

Variable	Hazard Ratio	95% Hazard Ratio Confidence Limits	
GROUP	1.198	0.520	2.758

Analysis of Maximum Likelihood Estimates
with Sandwich Variance Estimate

Variable	DF	Parameter Estimate	Standard Error	StdErr Ratio	Chi-Square	Pr > ChiSq
GROUP	1	0.18036	**0.18106**	0.425	0.9924	0.3192

Analysis of Maximum Likelihood Estimates
with Sandwich Variance Estimate

Variable	Hazard Ratio		
GROUP	1.198	0.840	1.708

Another way to model the clustered survival data is through the introduction of a common random effect called frailty (Vaupel, 1979) into the Cox's proportional hazards model. The survival model with shared random effects (either environmental or genetic) is called shared frailty or random effects model. This model assumes that the observations within the same cluster share a common unobservable random covariate, which acts multiplicatively on the hazard rate of each unit in the same cluster. For computational convenience, the frailties are usually assumed to follow a gamma distribution, and the model is called shared gamma frailty model (Clayton, 1985).

The frailty models are already made available in STATA and Splus, yet have not been built into SAS. Based on the 2006 SASware Ballot Results, SAS Inc. is considering to add frailty models into SAS. However, there are several user developed macros including %GAMFRAIL for gamma frailty model, and %PS_FRAIL for the positive stable frailty model (Shu, and Klein, 1997), %SPGAM1 and %PGAM2 for the semiparametric and parametric conditional shared Gamma model (Vu, 2002).

The analysis of time to first ≥3-lines data using % GAMFRAIL is shown in Figure 20. Surprisingly, the standard error estimated from gamma frailty model is 0.410, which is much larger than robust sandwich estimate (0.181), and only slightly smaller than that ignoring the correlation (0.427).

Fig. 20 SAS Output from Gamma Frailty Model

```
                Independence Model
          Estimate       S.E.       Wald    P-value

 Group    0.14296    0.40929    0.12199    0.72688


           Testing Null Hypothesis: Frailty=0
                      Criterion
          No Frailty   Frailty Chi-Square   P-value

 -2 Log L   170.9542  148.7385    22.2157    0.0000

                   Frailty Model
          Estimate       S.E.       Wald     Pvalue

 Frailty   9.13760    4.50000    4.12325    0.04230
 Group     0.16705    0.40967    0.16628    0.68344
```

CONCLUSIONS

The SAS is very powerful for analyzing clustered data. As summarized in Table 1, there are usually several different SAS procedures available to choose from for the analysis of a specific data. Which procedure to use really depends on the nature of the data and the parameters of interest. When the fixed regression parameters are of primary interest and the correlation structure is merely a nuisance, the marginal GEE models can be invaluable. On the other hand, GEE models are of limited use if the correlation structure is of primary interest, in such case, the mixed models or generalized linear mixed models will be more appropriate. Although in most of the situations, analysis using different SAS procedures yields similar results, it is recommended that the user should perform a sensitivity analysis using different available procedures and compare their results, especially when the p-values from one procedure is around the significant level.

With the development of new statistical methodology and its wide application, the SAS may consider to develop new SAS procedures, such as frailty models for clustered survival data, the zero-inflated model for the Poisson data, and generalized linear and latent model for the data with multilevel correlations, most of these models are already available in either STAT or Splus.

Table 1 The Summary for SAS Procedures for the analysis of clustered data

	Statistical Method	SAS Procedure
Continuous Data		
Non-model based	Paired ttest	PROC TTEST
	Signed rank test	PROC UNIVARIATE
	Modified Wilcoxon Rank test	%CLUSWILCOX
Model-based	Generalized Estimating Equations	PROC GENMOD
	Mixed model	PROC MIXED
	Generalized linear mixed model	PROC GLIMMIX
Binary Data		
Non-model based	McNemar's test	PROC FREQ
	Extended McNemar's test	%CLUSTPRO
	Modified Mantel-Hanszel test	%MHADJUST
Model-based	Generalized Estimating Equations	PROC GENMOD
	Generalized linear mixed model	PROC GLIMMIX
	Nonlinear mixed model	PROC NLMIXED
	Conditional logistic regression	PROC LOGISTIC with STRATA statement
Count Data		
Model-based	Generalized Estimating Equations	PROC GENMOD
	Generalized linear mixed model	PROC GLIMMIX

	Nonlinear mixed model	PROC NLMIXED
	Zero inflated Poisson Model	PROC TRAJ
Ordinal Data		
Model-based	Generalized Estimating Equations	PROC GENMOD
	Generalized linear mixed model	PROC GLIMMIX
	Bivariate Dale Model	%BDM
Time to Event Data		
Model-based	Marginal Model	PROC PHREG
	Frailty models	%GAMFRAIL, %PS_FRAIL %SPGAM1, %PGAM2

REFERENCES

Begg MD. Analyzing k (2 x 2) Tables under cluster sampling. *Biometrics* 1999; 55: 302-307.

Begg MD, Paykin AB. Performance of and software for a modified Mantel-Haenszel statistic for correlated data. J. *Statist. Comput. Simul.* 2001; 70:175-179.

Clayton DG, Cuzick J. Multivariate generalizations of the proportional hazards model (with discussion). *Journal of the Royal Statistical Society, Series B*. 1985; 148: 82-117.

Dale J. Global cross-ratio models for bivariate, discrete, ordered response. *Biometrics* 1986; 42:909-917.

Duff SW, Rohan TE, Altman DG. A method for combining matched and unmatched binary data. Application to randomized, controlled trials of photocoagulation in the treatment of diabetic retinopathy. *Am J Epidemiol* 1989; 130:371-8.

Johnston G, Strokes M. Application of GEE methodology using the SAS system. *NESUG* 1997.

Jones BL, Nagin DS, Roeder K. A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods & Research* 2001; 29:374-393.

Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; 34:1-14.

Liang KY, Zeger SL. Longitudianl data analysis using generalized linear models. *Biometrika* 1986; 73: 13-22.

Lieber ML, Ashley C. A SAS Macro implementing an extension of McNemar's Test for Clustered Data. *SUGI* 23. Paper 204, 1998.

Littell RC, Milliken GA, Stroup WW, Wolfinger RD. *SAS*® *System for Mixed Models*, Cary, NC: SAS Institute Inc., 1996. 633 pp.

McMillam G, Hanson T. SAS macro BDM for fitting the Dale regression model to bivariate ordinal response data. *Journal of Statistical Software* 2005; 14:1-12.

Obuchowski NA. On the Comparison of Correlated Proportions for Clustered Data. *Statistics in Medicine*, 1998; 17: 1495-1507.

Rosner B, Glynn RJ, Lee ML. Incorporation of clustering effects for the Wilcoxon rank sum test: a large-sample approach. *Biometrics* 2003; 59:1089-98.

Shu Y, Klein J. A SAS Macor for the positive stable frailty model. Master Thesis 1997; Medical College of Wisconsin, Milwaukee, Wisconsin.

The Choroidal Neovascularization Prevention Trial Research Group. Choroidal Neovascularization in the Chor-

oidal Neovascularization Prevention Trial. *Ophthalmology* 1998; 105:1364-1372.

The GLIMMIX Procedure. Cary, NC: SAS Institute Inc., November, 2005, 256 pp.

Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 1979; 16:439-454.

Vu HT. SAS Macors for parametric and semiparametric conditional shared gamma and log-normal frailty models. *Computation Statistics and Data Analysis* 2002; 40:173-187.

Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by using the marginal distributions. *Journal of American Statistical Association* 1989; 84:1065-1073.

ACKNOWLEDGMENTS

We would like to thank Maureen Maguire, Ph.D. for her support in writing this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Gui-shuang Ying, Ph.D.
Center for Preventive Ophthalmology and Biostatistics
3535 Market Street, Suite 700
Philadelphia, PA 19104
Work Phone: 215-615-1514
Fax: 215-615-1531
Email: gsying@mail.med.upenn.edu